Final Project Report: Crime Rate and Socioeconomic Factors

Student Information

- **Course:** Introduction to Data Science (DSA210)
- **Student:** Ali Murat Gültekin

Project Title

Analysis of Crime Rate and Socioeconomic Factors Using Data Science Techniques

---

Project Objective

This project aims to explore the relationship between crime rates (violent and property crimes) and various socioeconomic indicators (education, unemployment) across U.S. specific cities and states. The goal is to identify possible predictive links and provide data driven insights that could support crime prevention and policymaking.

Dataset Description

1. Crime Data:

- Source: FBI Crime Data Explorer (2023)
- Scope: City-level violent and property crime statistics

2. Socioeconomic Data:

- Source: USDA Economic Research Service (2023)
- Variables: Percentage of population with a bachelor’s degree or higher, unemployment rate

3. Combined Dataset:

- Final dataset created by merging crime data with socioeconomic indicators
- Size: near 18,000 records (cities/states)

Exploratory Data Analysis (EDA)

- **Heatmaps**: Correlation between crime types and education/unemployment

- **Scatter Plots**:
  - Violent crime vs. education → no clear trend
  - Property crime vs. unemployment → slight upward trend
- **Descriptive Stats**:
  - Avg. education: ~26.6%
  - Avg. unemployment: ~3.8%

Hypothesis Testing

1. **H□:** No correlation between education and violent crime
   - Pearson r = 0.0041 → no linear relationship
   - Spearman ρ = 0.0406 ($p < 0.001$) → negligible
2. **H□:** No correlation between unemployment and property crime
   - Pearson r = 0.0537 ($p < 0.001$)
   - Spearman ρ = 0.1453 ($p < 0.001$) → weak but significant

---

Machine Learning Modeling

- **Objective**: Classify cities as "high crime" or "low crime"
- **Target Variables**:
  - Binary labels for violent and property crime (based on median split)
- **Features**: Education %, Unemployment Rate

Models Applied:

1. Logistic Regression
2. Random Forest Classifier
3. K-Nearest Neighbors (KNN)

Evaluation Metrics:

- Accuracy
- Precision
- Recall
- F1 Score

Final Results & Interpretation

In the final phase of the project, machine learning classification models were applied to predict whether a city has a **high or low level of crime**, using only **education level** and **unemployment rate** as input features. Two binary classification tasks were conducted:

- violent_crime_class: High vs. Low violent crime
- property_crime_class: High vs. Low property crime

Model Performance Summary

Target	Model	Accuracy	Precision	Recall	F1 Score
Violent Crime	Logistic Regression	0.64	0.62	0.68	0.65
Violent Crime	Random Forest	0.71	0.70	0.73	0.71
Violent Crime	K-Nearest Neighbors	0.70	0.69	0.71	0.70
Property Crime	Logistic Regression	0.63	0.61	0.66	0.63
Property Crime	Random Forest	0.69	0.68	0.71	0.69
Property Crime	K-Nearest Neighbors	0.67	0.65	0.68	0.66

Extra Note: Values above reflect summarized results from final notebook. Slight variations may happen depending on preprocessing or random seed values.

Key Findings

- **Random Forest** consistently achieved the best performance across both tasks.
- **Unemployment** had slightly higher predictive power than education.
- **Logistic Regression** coefficients confirmed weak but interpretable trends.
- Socioeconomic variables alone provide limited but measurable predictive ability for crime classification.

Tools & Technologies

- Python
- Jupyter Notebook

- Libraries: pandas, scikit-learn, matplotlib, seaborn

Ethical Considerations

- All data used is anonymized and publicly available
- Bias in crime reporting acknowledged
- Transparency in data cleaning and merging procedures ensured

---

Final Deliverables

- Jupyter Notebooks (EDA, Hypothesis Testing, ML Modeling)
- README.md (contains project phases and instructions)
- requirements.txt for dependencies

Conclusion

This study showed that while education had minimal correlation with violent crime, unemployment had a small but statistically significant relationship with property crime. Machine learning models confirmed that socioeconomic data alone offers limited but useful predictive power for understanding crime distributions. More comprehensive models may require additional features like income, housing conditions, or law enforcement data.

Data with visuals(graphs, curve, …) can be accessed from README.md in repo.